Bayesian Averaging over Decision Tree Models: an Application for Estimating Uncertainty in Trauma Severity Scoring

V. Schetinin*, L. Jakaite

University of Bedfordshire

W. Krzanowski

University of Exeter

Abstract

Introduction: For making reliable decisions, practitioners need to estimate uncertainties that exist in data and decision models. In this paper we analyse uncertainties of predicting survival probability for patients in trauma care. The existing prediction methodology employs logistic regression modelling of Trauma and Injury Severity Score (TRISS), which is based on theoretical assumptions. These assumptions limit the capability of TRISS methodology to provide accurate and reliable predictions.

Methods: We adopt the methodology of Bayesian model averaging and show how this methodology can be applied to Decision Trees in order to provide practitioners with new insights into the uncertainty. The proposed method has been validated on a large set of 447,176 cases registered in the US National Trauma Data Bank in terms of discrimination ability evaluated with Receiver Operating Characteristic (ROC) and Precision-Recall (PRC) Curves.

Results: Areas under curves were improved for ROC from 0.951 to 0.956 ($p = 3.89 \cdot 10^{-18}$) and for PRC from 0.564 to 0.605 ($p = 3.89 \cdot 10^{-18}$). The new model has significantly better calibration in terms of the Hosmer-Lemeshow $\hat{H}$ statistic, showing an improvement from 223.14 (the standard method) to 11.59 ($p = 2.31 \cdot 10^{-18}$).

Conclusion: The proposed Bayesian method is capable of improving the accuracy and reliability of survival prediction. The new method has been made available for evaluation purposes as a web application.

Keywords: Bayesian Model Averaging, Decision Tree, Predictive posterior distribution, Trauma and Injury Severity Scoring, TRISS

1. Introduction

Decision making in health care is subject to uncertainties that exist in data and decision models. In this regard Machine Learning (ML) methods have been intensively developed over the last decade to provide practitioners with reliable estimates of uncertainties in decisions and predictions, see e.g. [1]. Using ML methods, predictions of functional recovery and mortality after traumatic brain injury were considered in [2, 3]. The combination of Glasgow Coma Scale scores with clinical and laboratory parameters of patients has provided a high prediction accuracy. Prediction of burn patient survival was undertaken in [4] using models that were developed on patient data. The data included information about the patient's age, sex, and percentage of burns in eight parts of the body.

In trauma care the evaluation of injury severity of patients has a long history of using logistic regression modelling known as the Trauma and Injury Severity Score (TRISS) [5, 6, 7, 8]. The TRISS model allows practitioners to predict the probability of survival for a patient on arrival at a hospital. The predictions are based on screening tests which are recorded at an accident scene. Screening tests are evaluated by a trained scorer for injuries which a patient can obtain in the following six regions of the body: head, face, chest, abdomen, extremities, and external (skin, subcutaneous tissue and burns).

Estimates of survival probabilities of patients enable practitioners to identify cases for peer review and compare the survival rates of different patient groups. TRISS estimates are also used for benchmarking and monitoring of patient outcomes over time and between hospitals [9, 10].

Uncertainties that exist in data as well as in the prediction model affect the results and might lead to misleading decisions. For this reason, practitioners have raised a concern about the ability of TRISS to provide reliable predictions and estimates of uncertainty [9, 6, 11].

The accuracy of predictions is compared against actual survival during development of prediction models. The relationship between predicted and actual probabilities can be visualised as a calibration curve [12, 13]. Trauma care practitioners have found that the TRISS calibration curve is not ideal [14, 7, 11].

It has been found that the accuracy of TRISS predictions is acceptable when the types and severities of patient injuries are typical [6]. However, for patients with four or more injuries as well as those with atypical combinations of injuries, the accuracy has to be improved. In practice, it is critically important to reliably estimate the uncertainty in a predicted survival probability. The uncertainty estimates are required in order to minimise risks of misleading decisions. Uncertainty can be represented by confidence intervals. These intervals are reliably estimated when the density of predicted probabilities is fully tractable, which is achievable only in trivial cases. Thus TRISS methodology that is based on theoretical assumptions cannot realistically estimate the uncertainty.

*Corresponding author
Email address: v.schetinin@beds.ac.uk (V. Schetinin)

To tackle such problems, we adopt the methodology of Bayesian learning of models, which in theory provides reliable estimation of uncertainty intervals [15, 16, 17]. This approach, however, requires intensive computations, as discussed in [18, 19].

The learning methodology is based on Bayesian Model Averaging (BMA) which defines a prediction model with parameters that determine the prediction ability. We use the Bayesian method for averaging over Decision Tree (DT) models which are known for their ability to select input variables that are relevant to the problem, as discussed in [20, 21, 22]. The given data are recursively split along input variables into reasonably small subsets. Splits made along variables are easy-to-interpret, and thus DT models built on the given data are able to assist practitioners with new insights [23].

In most practical cases, any given model is incapable of fully explaining the real-world data, which means that a single "true" model does not exist. The BMA methodology, adopted in our study, assumes that different models can be mixed together so that their average under certain conditions will approximate the true model of interest. The averaging strategy is often more efficient than model selection in real-world applications when the predictive ability (or fitness function) is not unimodal [24, 25].

The trauma injury severity scoring problems have been considered in the Bayesian context. The study described in [26, 27, 28] has been undertaken with the main focus on the Receiver Operating Characteristic (ROC) curve [29], following the standard practice in diagnostic test evaluation. This evaluation, however, is insufficient in the case of imbalanced patient data with a low rate of positive outcomes (e.g. mortality), see e.g. [30, 31].

In this paper we propose a new approach to estimating the predictive posterior probability densities of injured patients and examine the accuracy and reliability of predictions on the patient data with low mortality rate. The proposed method is examined on the cases which were registered in the US National Trauma Data Bank (NTDB) [32] with multiple injuries. We also describe a web application [33] which has been developed and made available for evaluation by trauma care practitioners. The application assists practitioners to deliver reliable estimates of uncertainty intervals within which predicted survival probabilities are expected for a patient. Finally we discuss the main results, and draw conclusions.

2. Material and methods

2.1. Data

Our study is conducted on cases from the US NTDB, the major source of data about injured patients admitted to hospitals and emergency units [32]. The data include patient age, gender, type and regions of injuries along with some clinical and background information about patient state. The NTDB also includes the TRISS prediction and the outcome of care, alive or died, for each patient.

Table 1: Screening tests.

#	*Notation*	*Name*	*min*	*max*
1	A	Age	0	100
2	G	Gender	0 female	1 male
3	T	Injury type	0 penetrating	1 blunt
4	B	Blood pressure	0	300
5	R	Respiration rate	0	200
6	G_E	GCS Eye	1	4
7	G_V	GCS Verbal	1	5
8	G_M	GCS Motor	1	6
9	H_S	Head severity	0	6
10	F_S	Face severity	0	4
11	N_S	Neck severity	0	6
12	T_S	Thorax severity	0	6
13	A_S	Abdomen severity	0	6
14	S_S	Spine severity	0	6
15	U_S	Upper extremity severity	0	6
16	L_S	Lower extremity severity	0	6
17	E_S	External severity	0	6

Table 1 shows the screening tests (or predictors) that are used by the TRISS method. The variables *Age*, *Blood pressure*, and *Respiration rate* are continuous, and the remaining variables are ordinal. The patient outcome is the *discharge status*: 0 is alive, and 1 is died. The table also shows the minimal and maximal values of each test.

For our study, we selected patient records which do not contain missing values in the above predictors. The maximal number of injuries in these records was 48. The number of these cases was 571,775, including 384,876 cases with 1-3 injuries and 186,899 with 4-48 injures. As we discussed in Section 1, the TRISS model has a limited ability to predict outcomes of patients with more than three injuries.

The injuries in NTDB are recorded using to the Abbreviated Injury Scale (AIS) codes, which assign a severity score to each injury. The AIS severity scores are based on mortality risk and range from 1, minor injuries including wounds to skin or subcutaneous tissue or closed fractures [34], to six, considered fatal [35]. However, it has been recently found that about 48.6% of patients with an injury severity of 6 survive [36].

Table 2 shows statistics of the screening tests A to E_S listed in Table 1 in the above groups of patients. The statistics are represented by values of the mean, standard deviation, median, and quartiles.

Table 2: Statistics of the screening tests over four patient groups with the following number of injuries: 1-48,1-3,4-20, and 21-48

	Injuries	A	G	T	B	R	G_E	G_V	G_M	H_S	F_S	N_S	T_S	A_S	S_S	U_S	L_S	E_S
Mean	1-48	39.3	0.7	0.9	134.5	19.1	3.7	4.5	5.6	0.9	0.4	0.0	0.7	0.4	0.4	0.6	0.8	0.1
	1-3	39.8	0.6	0.9	136.1	19.3	3.8	4.7	5.8	0.6	0.2	0.0	0.3	0.2	0.2	0.4	0.7	0.1
	4-48	38.2	0.7	0.9	131.3	18.7	3.5	4.2	5.2	1.6	0.7	0.1	1.4	0.7	0.7	0.9	1.1	0.1
Standard deviation	1-48	22.6	0.5	0.3	29.7	7.1	0.8	1.1	1.3	1.5	0.7	0.3	1.3	1.0	0.9	0.9	1.2	0.3
	1-3	23.7	0.5	0.3	28.5	6.0	0.6	0.9	1.0	1.2	0.5	0.2	0.9	0.8	0.7	0.8	1.1	0.3
	4-48	20.0	0.5	0.3	31.8	8.9	1.1	1.5	1.7	1.7	0.8	0.3	1.6	1.2	1.1	1.0	1.2	0.4
Median	1-48	36	1	1	135	20	4	5	6	0	0	0	0	0	0	0	0	0
	1-3	36	1	1	136	19	4	5	6	0	0	0	0	0	0	0	0	0
	4-48	35	1	1	133	20	4	5	6	1	0	0	0	0	0	1	1	0
1^{st} quartile	1-48	21	0	1	119	16	4	5	6	0	0	0	0	0	0	0	0	0
	1-3	21	0	1	120	16	4	5	6	0	0	0	0	0	0	0	0	0
	4-48	22	0	1	116	16	4	4	6	0	0	0	0	0	0	0	0	0
3^{rd} quartile	1-48	54	1	1	150	22	4	5	6	2	1	0	1	0	0	1	2	0
	1-3	55	1	1	151	20	4	5	6	0	0	0	0	0	0	0	2	0
	4-48	51	1	1	149	22	4	5	6	3	1	0	3	1	2	2	2	0

2.2. Logistic regression model

The use of logistic regression is a conventional way of predicting survival probabilities [5, 13]. In trauma injury severity scoring, the logistic regression includes screening tests which are ordinal and continuous predictors. The ordinal variables represent severity scores of injuries obtained by a patient, the Glasgow Coma Scale (GCS), and the injury type. The continuous variables include age, systolic blood pressure, and respiratory rate.

The TRISS model is based on the above screening tests in Table 1, which are represented by the following two aggregated scores: Injury Severity Score (ISS), and Revised Trauma Score (RTS) [5, 37]. A side effect of using the aggregated scores is unexplained fluctuations in the calculated survival probabilities, which affect the prediction accuracy, as discussed in [6, 38].

The TRISS model determines the probability of survival, P, in the following logistic regression form:

$$P = \frac{1}{1 + e^{-b}}, \tag{1}$$

where $b = b_0 + b_1 \times RTS + b_2 \times ISS + b_3 \times A$.

Here $b_0, \ldots, b_3$ are the regression parameters, and A is the dichotomised age: $A = 0$, if $age < 55$, and $A = 1$, otherwise. The parameters b were separately determined for blunt and penetrating types of injuries. As discussed in Section 1, the above TRISS model can consider only up to three of the most severe injuries which a patient can obtain.

The TRISS model assumes that the density of predicted values is Gaussian, $N(\mu, \sigma^2)$, where μ and σ^2 are the mean and standard deviation. This assumption, however, is often unrealistic, as discussed in [39, 40, 41]. This issue has been examined in the previous work [42] in which we found that the calculated uncertainty intervals can be biased.

The above TRISS model, described by Eq. 1, has been made available online for calculating survival probabilities [43]. The TRISS calculator allows practitioners to calculate the survival probability for a given patient test.

2.3. Bayesian model averaging

In practice, the probability distributions of data and model parameters cannot be specified so as to meet the full requirements of Bayesian methodology. Except for trivial cases, the BMA can be practically implemented with Markov chain Monte Carlo (MCMC) methods, as described in [24]. Bayesian averaging over DT models can be also efficiently implemented with MCMC [44, 18]. In our previous study, however, we have found that the MCMC can make excessive samples of oversized DT models, which degrade the approximation, as discussed in [42, 45, 46].

To outline the MCMC approximation let a predictive model have parameter vector Θ, input vectors $x = (x_1, \ldots, x_m)$, and outcomes y, where m is the dimensionality of x. The training data $\mathbf{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{n}$ include n instances. According to this notation an input vector x is assigned to one of the given

classes, $y \in \{1, C\}$. In our case, there are $C = 2$ classes, $y = 0$ if a patient survived, and $y = 1$ if died.

The predictive posterior distribution of interest, $p(y|x, \mathbf{D})$, is calculated as an integral over model parameters Θ as follows:

$$p(y|x, \mathbf{D}) = \int_{\Theta} p(y|x, \Theta) p(\Theta|\mathbf{D}) \mathrm{d}\Theta, \tag{2}$$

where $p(y|x, \Theta)$ is the posterior predictive density given input x and model parameters Θ, and $p(\Theta|\mathbf{D})$ is the posterior density of Θ given data $\mathbf{D}$.

The MCMC algorithm generates N samples, $\left\{\Theta^{(i)}\right\}_{i=1}^{N}$, which are distributed with a density function $\hat{p}(\Theta|\mathbf{D})$:

$$\Theta^{(i)} \sim \hat{p}(\Theta|\mathbf{D}). \tag{3}$$

The desired approximation is achieved when the MCMC algorithm generates a random sequence with a stationary probability distribution. Thus we can draw samples $\Theta^{(i)}$ defined in Eq. 3 and then calculate the predictive density of interest as follows:

$$p(y|x, \mathbf{D}) \approx \sum_{i=1}^{N} p(y|x, \Theta^{(i)}, \mathbf{D}) p(\Theta^{(i)}|\mathbf{D}) = \frac{1}{N} \sum_{i=1}^{N} p(y|x, \Theta^{(i)}, \mathbf{D}). \tag{4}$$

From Eq. 3, the required model parameters $\Theta^{(i)}$ are drawn from a posterior distribution simulated by MCMC. The collected samples are then used in Eq. 4 to calculate the posterior predictive probabilities. Details of the MCMC algorithm for sampling DT models are given in the Appendix.

2.4. Validation of prediction model

In our approach to developing a new prediction model we use two groups of patients registered in the NTDB without missing values described in Section 2.1 above. Group 1 includes 186,899 records with four and more injuries, mortality in which was 7.78%. The cases of this group are used for developing a model within a 3-fold cross-validation. Group 2 includes 384,876 patients with three and fewer injuries, mortality in which was 2.7%. The model which is developed on 124,598 cases from the first group is then validated on 447,176 cases of both groups. The average mortality of patients included in the validation data was 3.38%. Table 3 shows how the data are used within the 3-fold cross-validation. Here V and D stand for validation and development of model, respectively. The table shows that at each fold a model is developed on 124,598 cases of the Group 1 and then validated on a mix of 62,300 cases of Group 1 and all cases of Group 2.

Diagnostic and discrimination ability of models is evaluated in terms of AUC, the Area Under the Receiver Operating Characteristic (ROC), curve, which is a summary measure of the accuracy of a quantitative diagnostic test, see e.g. [29, 40]. For cases with a low rate of positive cases, $P(Y = 1)$, the use of the Precision-Recall Curve (PRC) is more informative than AUC. PRC shows a rate

Table 3: Use of patient groups within 3-fold cross-validation. D and V denote subsets for development and validation of model, respectively.

	Group 1			Group 2
# fold	1	2	3	
1	V	D	D	V
2	D	V	D	V
3	D	D	V	V

of true positive cases among predicted positive, $P(\hat{Y} = 1|Y = 1)$, versus a rate of true positive cases among positive cases, $P(Y = 1|\hat{Y} = 1)$, where $P(\hat{Y} = 1)$ are the rate of positive predictions. PRC thus reflects the ability of a decision model to identify positive cases for a given rate of positive predictions [30, 31].

The accuracy of predictions which are made by a model is evaluated by goodness-of-fit tests on the validation data. When cases have binary outcomes, $Y \in \{0, 1\}$, goodness-of-fit is estimated as an agreement between observed outcomes (mortality rate) and predicted probabilities. A plot which shows predicted probabilities along x-axis and the mortality rate along y-axis is defined as a calibration curve. The ideal predictions lie on the 45° line. The calibration curve shows the observed mortality rates for cases which are grouped by values of predicted probabilities, see e.g. [40].

For trauma survival prediction, calibration is typically evaluated by the Hosmer-Lemeshow (HL) statistic which is normally calculated for 10 intervals (deciles) of predicted values [12]. Under certain conditions the larger the HL statistic, the worse is the calibration. The HL-test, however, is dependent on sample size and becomes statistically significant when the number of cases exceeds 10,000 [47, 48]. Therefore, the HL-test has to be analysed along with the overall number of cases and results of other tests. In our experiments with large patient groups, the HL statistic was significant ($p = 10^{-8}$) and thus we use sensitivity and specificity along with Brier score to provide additional evidence, as discussed in [40].

There are two ways to arrange patients by values of predicted probabilities, which are namely $\hat{C}$ and $\hat{H}$-statistics [12]. The $\hat{C}$-statistic divides patients into groups with an equal number of cases arranged in ascending order of predicted probabilities. The predicted mortality within each group is determined by the cases in the group. In contrast, the $\hat{H}$-statistic forms groups of cases with predicted probabilities lying within equal ranges, making the numbers of cases in each group variable.

It has been shown that when a rate of positive outcomes is less than 5%, the $\hat{C}$ statistic fails to detect significant deviations of observed and predicted values, whilst the $\hat{H}$-statistic can detect such deviations [49]. For this reason in our experiments we use the $\hat{H}$-statistic.

We use a simulation (bootstrap) strategy for evaluating significance of tests, which is based on applying a test to a smaller set of random samples of the original data [50, 48]. Such a strategy has been recently used for evaluating the

HL-statistic of calibration for a trauma survival prediction model, described in [51]. The Brier score has been also tested within the similar simulation strategy described in [52, 41, 53].

The above simulation technique is adopted for our experiments. The experiments as well as the proposed method were implemented in MATLAB. The next section describes the results of the experiments.

3. Results

In this section we describe the main results which were obtained by the proposed and TRISS methods on the NTDB benchmark outlined in Section 2.1. The results are compared in terms of the PRC, AUC, HL-statistic and Brier score, as discussed in Section 2.4.

3.1. Calibration curves

According to the HL test discussed in Section 2, the calibration curves were calculated by using the HL $\hat{H}$-statistic for intervals which are equidistantly distributed over survival probabilities. Fig. 1 shows the calibration curves for the TRISS (Blue line) and BDT (Red line) models. The curves were calculated for 10 intervals (the left side plot) as well as for 20 intervals (the right side plot) to estimate the influence of interval ranges. We can observe that the BDT model is significantly better fitted to the ideal calibration shown as the 45°dashed line in both cases of 10 and 20 intervals. Table 4 shows the significant ($p = 2.31 \cdot 10^{-18}$) improvement of calibration, which is evaluated by the $\hat{H}$-statistic and Brier score. The significance was estimated by simulation according to the methodology described in Section 2. Note that the smaller the statistic or score, the better the calibration.

3.2. Discrimination abilities of TRISS and BDT models

According to the adopted methodology, the proposed BDT and TRISS models were compared in terms of discrimination ability estimated by ROC and PRC. Table 5 shows the AUC values estimated for ROC and PRC calculated for the TRISS and BDT models. The simulation of the AUC characteristics shows that the discrimination ability of the BDT model is significantly ($p = 3.9 \cdot 10^{-18}$) higher than that provided by the TRISS model.

Fig. 2 shows both the ROC and PRC calculated for the TRISS and BDT models. The plotted curves show areas where the BDT model outperforms the TRISS model in terms of recall, precision, or specificity.

Table 4: Significance of $\hat{H}$-statistic and Brier scores for TRISS and BDT methods.

	TRISS	BDT	p-value
$\hat{H}$-statistic	223.14	11.59	$2.31 \cdot 10^{-18}$
Brier score	0.025	0.023	$3.89 \cdot 10^{-18}$

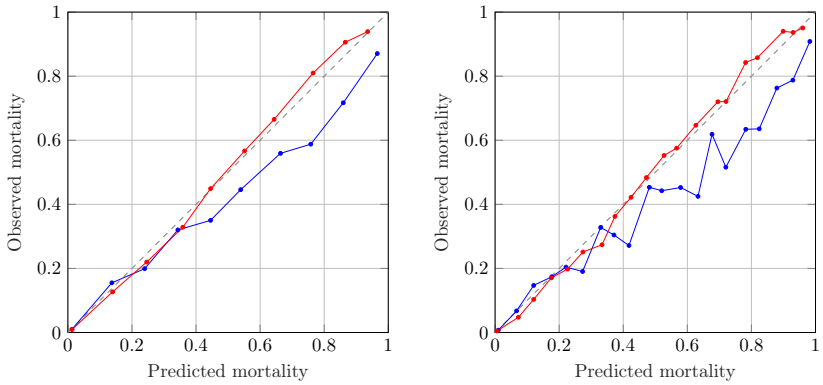

Figure 1: Calibration curves for TRISS (Blue line) and proposed BDT (Red line) models. The curves calculated for 10 (the left plot) and 20 (the right plot) intervals.

Table 5: Significance of AUC characteristics of ROC and PRC calculated for the TRISS and BDT models.

AUC	TRISS	BDT	p-value
ROC	0.951	0.956	$3.89 \cdot 10^{-18}$
PRC	0.564	0.605	$3.89 \cdot 10^{-18}$

Table 6 shows how the recall, precision, F_1, and specificity are dependent on thresholds Q. Here F_1 is the score showing a balance between precision and recall of a model. In particular, the BDT model for $Q = 0.375$ has the maximum F_1 value of 0.63, at which both the recall and precision are higher than those provided by the TRISS model. The indicators I_R and I_P included in Table 6 show the other areas of Q where the BDT model outperforms the TRISS model in terms of recall or precision. The areas where the BDT model has improved both these characteristics are shown in bold.

Table 6: Recall, Precision, F_1, and Specificity over thresholds Q for the TRISS (TRS) and BDT models.

			Recall		*Precision*		F_1		*Specificity*	
Q	I_R	I_P	*TRS*	*BDT*	*TRS*	*BDT*	*TRS*	*BDT*	*TRS*	*BDT*
0.025	0	1	0.918	0.931	0.153	0.149	0.262	0.257	0.821	0.813
0.050	0	1	0.820	0.889	0.296	0.195	0.435	0.320	0.932	0.870
0.075	0	1	0.768	0.842	0.394	0.257	0.521	0.393	0.958	0.914
0.100	0	1	0.753	0.794	0.432	0.341	0.549	0.477	0.965	0.946
0.125	0	1	0.706	0.754	0.483	0.406	0.574	0.527	0.973	0.961
0.150	0	1	0.692	0.733	0.510	0.438	0.587	0.548	0.977	0.967
0.175	0	1	0.684	0.712	0.527	0.473	0.595	0.568	0.978	0.972
0.200	0	1	0.664	0.686	0.551	0.526	0.602	0.595	0.981	0.978
0.225	0	1	0.642	0.671	0.579	0.550	0.609	0.605	0.984	0.981
0.250	0	1	0.631	0.648	0.595	0.586	0.613	0.616	0.985	0.984
0.275	1	1	0.625	**0.629**	0.606	**0.613**	0.615	**0.620**	0.986	0.986
0.300	1	0	0.618	0.615	0.621	0.631	0.620	0.623	0.987	0.987
0.325	1	1	0.606	**0.608**	0.637	**0.642**	0.621	**0.624**	0.988	0.988
0.350	1	1	0.570	**0.596**	0.659	**0.660**	0.611	**0.626**	0.990	0.989
0.375	1	1	0.558	**0.571**	0.672	**0.703**	0.610	**0.630**	0.990	0.991
0.400	1	0	0.554	0.548	0.679	0.728	0.610	0.624	0.991	0.993
0.425	1	0	0.545	0.531	0.698	0.744	0.612	0.619	0.992	0.994
0.450	1	0	0.540	0.517	0.707	0.759	0.612	0.615	0.992	0.994
0.475	1	0	0.532	0.499	0.710	0.785	0.609	0.610	0.992	0.995
0.500	1	0	0.511	0.488	0.726	0.796	0.599	0.604	0.993	0.996
0.525	1	0	0.475	0.474	0.741	0.809	0.579	0.597	0.994	0.996
0.550	1	1	0.461	**0.464**	0.754	**0.819**	0.573	**0.591**	0.995	0.996
0.575	1	0	0.452	0.419	0.761	0.861	0.567	0.563	0.995	0.998
0.600	1	0	0.441	0.402	0.771	0.880	0.561	0.552	0.995	0.998
0.625	1	0	0.437	0.387	0.775	0.892	0.559	0.540	0.996	0.998
0.650	1	0	0.424	0.373	0.792	0.904	0.552	0.528	0.996	0.999
0.675	1	0	0.398	0.373	0.798	0.904	0.531	0.528	0.996	0.999
0.700	1	1	0.359	**0.366**	0.812	**0.908**	0.498	**0.522**	0.997	0.999
0.725	1	1	0.347	**0.362**	0.825	**0.911**	0.489	**0.517**	0.997	0.999
0.750	1	1	0.341	**0.344**	0.834	**0.920**	0.484	**0.500**	0.998	0.999
0.775	1	0	0.335	0.329	0.838	0.923	0.478	0.485	0.998	0.999
0.800	1	1	0.297	**0.299**	0.862	**0.933**	0.442	**0.452**	0.998	0.999
0.825	1	0	0.286	0.261	0.869	0.944	0.431	0.408	0.998	0.999
0.850	1	0	0.274	0.261	0.881	0.944	0.418	0.408	0.999	0.999
0.875	1	1	0.256	**0.260**	0.888	**0.944**	0.398	**0.408**	0.999	0.999
0.900	1	1	0.205	**0.221**	0.898	**0.944**	0.334	**0.356**	0.999	1.000
0.925	1	1	0.190	**0.201**	0.908	**0.946**	0.315	**0.331**	0.999	1.000

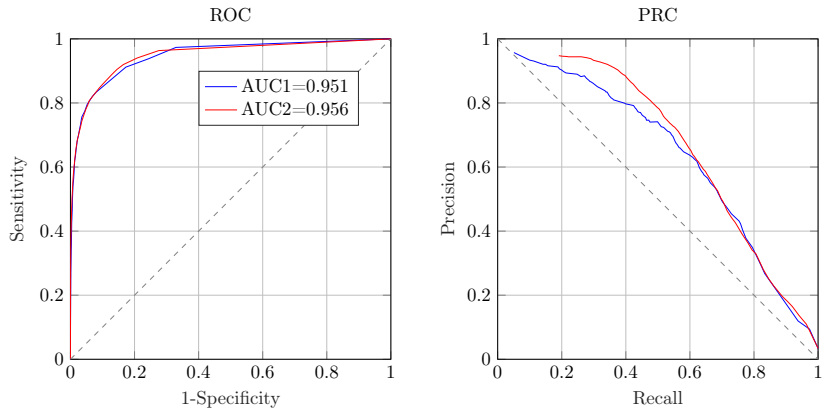

Figure 2: ROC (on the left side) and PRC (on the right side) for the TRISS (Blue line) and BDT (Red line) models. AUC1 and AUC2 are shown for the TRISS and BDT models, respectively.

3.3. Single DT model with MAP

Fig. 3 shows the Maximum A-*Posteriori* (MAP) DT that was found in the DT models collected for the Bayesian averaging. The DT model shows the recursive partitions of the labelled data set represented by the screening tests listed in Table 1. The DT model built on these data was pruned to 36 terminal nodes. Each splitting node in the DT model examines a screening test $x_i \in \{x_1, x_{17}\}$. The examination starts at the root node with a test, G_M (GCS Motor) and then it is continued into one of two branches, left or right, that connect the other nodes. The connected nodes are recursively examined until arrival at a terminal node that will finally assign a survival probability to the patient. For example, consider a case with the following test readings: $G_M = 4$ (GCS Motor), $T = 0$ (Injury type), and $B = 19$ (Blood pressure). For this case, the MAP DT predicts a low survival probability 0.01.

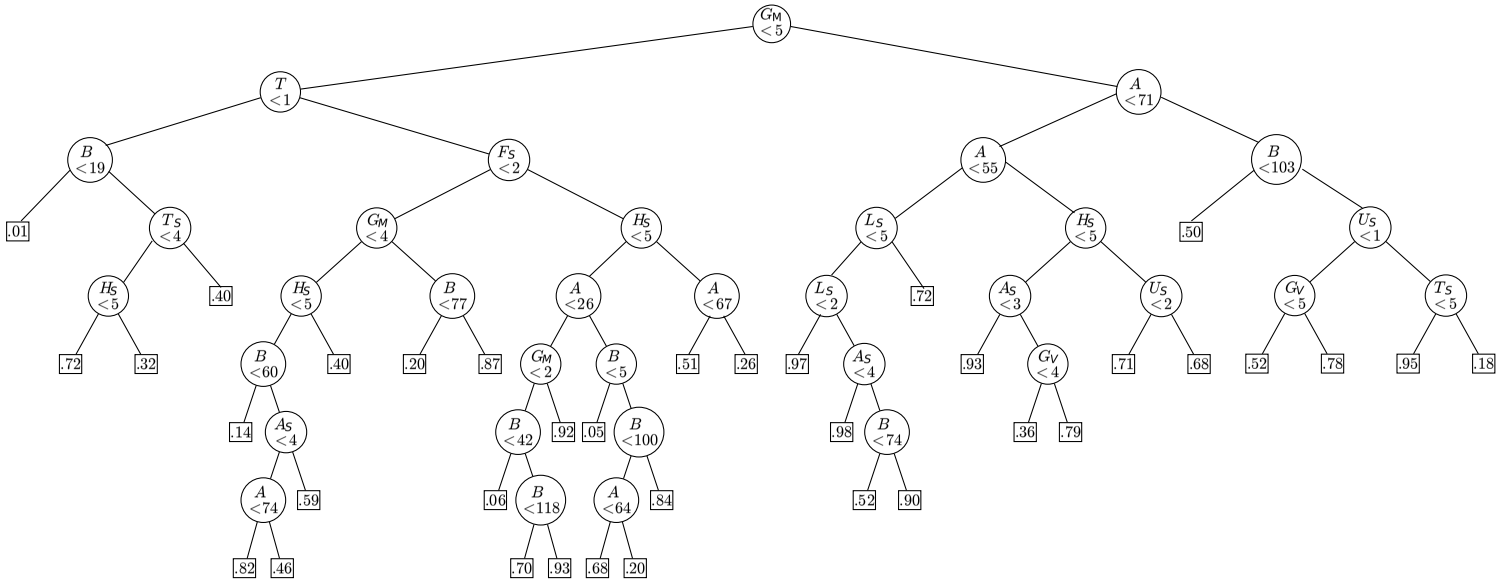

Figure 3: Maximum a-posteriori DT model. Legend: Age (A), Injury type (T), Blood pressure (B), GCS Verbal (G_V), GCS Motor (G_M), Head severity (H_S), Face severity (F_S), Thorax severity (T_S), Abdomen severity (A_S), Upper extremity severity (U_S), Lower extremity severity (L_S).

3.4. Web Application

The proposed method has been implemented as a web application `www.TraumaCalc.org` with an interface shown in Fig. 4. The left-hand interface panel shows fields from 1 to 17 with the patient's data filled in according to the screening tests in Table 1. The fields are shown with their names and permitted ranges. On the right-hand panel there is a plot of the *Predicted probability distribution* that displays a histogram calculated for survival probabilities of the given patient. The histogram shows an interval within which the probabilities are most frequently distributed. The interval represents the prediction uncertainty that is associated with risk of making a misleading decision. The larger the interval, the greater is the uncertainty and higher the risk for a patient. The estimate of the predictive density is delivered for the given tests of the patient's condition. The density (bars in Blue) is shown along with the average predicted probability. The predicted probability places the given patient within the highest density interval (0.65,0.95) indicating that this patient will most probably survive.

The button *Calculate* shown in the Fig. 4 initiates the calculation of the probability distribution for a given patient's data. The data are transferred onto the server to be processed by the Bayesian model described in Section 2.3. Because of heavy computations, a high performance implementation of this model was developed.

The application can be run in a web browser on a user's device. Intensive computations are made on the server, and so large computation power is not required on the user's side.

Finally it is important to note that the proposed Bayesian method delivers estimates of uncertainty intervals which are individual for a patient. The TRISS method can only provide estimates of uncertainty intervals, which have been calculated on a patient population selected for developing the prediction model, reflecting the variability of predicted outcomes in the given population.

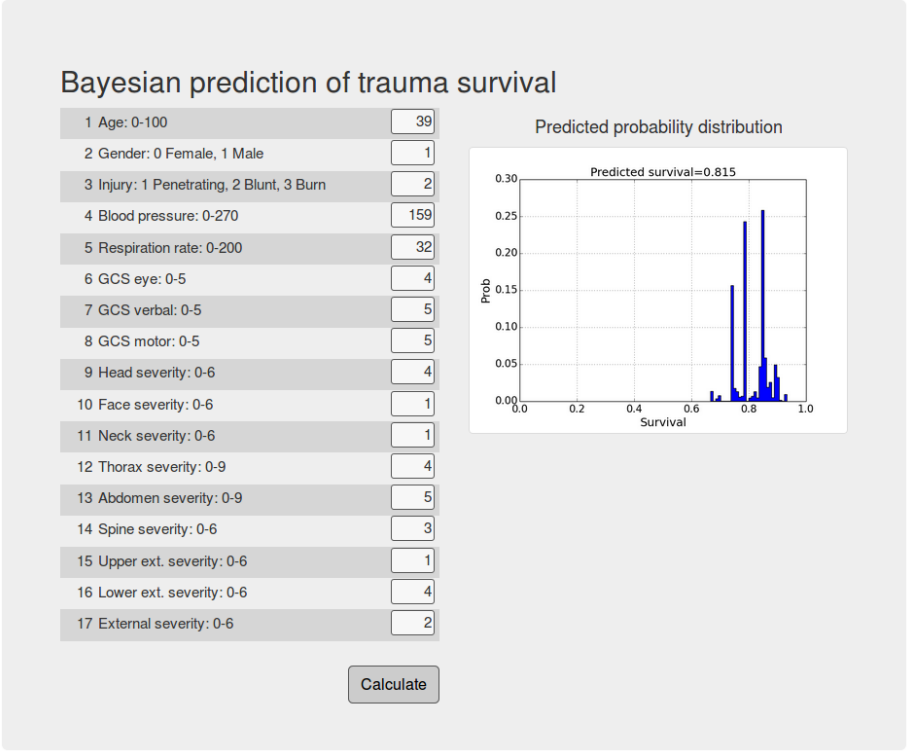

Figure 4: Web-based application for estimating the predictive probability distribution of survival.

4. Discussion and Conclusion

4.1. Contribution

TRISS is the current standard for evaluating injury severity and predicting outcomes of patients on arrival at a hospital. The screening tests conducted by a scorer are analysed by the TRISS method based on logistic regression described in Section 2.2.

The TRISS evaluations are affected by uncertainties that exist in both data and prediction model. These uncertainties increase risks of making misleading decisions. Practitioners may not be satisfied with the ability of TRISS to make reliable predictions when errors in the tests affect patient outcomes. They may be concerned that the goodness-of-fit of the TRISS model is not ideal. It has also been suggested that the accuracy of TRISS predictions for patients with four or more injuries as well as with atypical injuries needs to be improved [11, 6].

It is known that uncertainty intervals within which predictions are distributed can be reliably estimated in an analytical form if the distribution function of predicted probabilities is known. However, the TRISS methodology is based on theoretical assumptions about probability distributions, and so cannot provide reliable estimates of the uncertainty intervals [6].

To address the above problems we have proposed a Bayesian approach. Bayesian model averaging is well known for reliable modelling and estimation of uncertainty, which however require intensive computations. The methodology applied to DT models is made feasible with MCMC, which under certain conditions can accurately approximate a probability distribution of interest [24, 18]. We developed a new model which in our experiments was validated on the US NTDB, as described in Section 3. We found that the new model has significantly better calibration in terms of the Hosmer-Lemeshow $\hat{H}$ statistic, showing an improvement from 223.14 (TRISS model) to 11.59 ($p = 2.31 \cdot 10^{-18}$). Moreover, the Brier score, which is also used for evaluating goodness-of-fit of models, was improved from 0.025 to 0.023 ($p = 3.89 \cdot 10^{-18}$).

The new model has outperformed TRISS in terms of discrimination ability evaluated with ROC and PRC. Areas under curves were improved for ROC from 0.951 to 0.956 ($p = 3.89 \cdot 10^{-18}$) and for PRC from 0.564 to 0.605 ($p = 3.89 \cdot 10^{-18}$). We found that the new model has outperformed TRISS in terms of precision and recall providing a higher F_1 score, 0.63.

In summary, our contributions are as follows.

(1) It is evident that the accuracy of the TRISS model has to be improved for patients who obtained four and more injuries, and so have a high risk of death. The TRISS methodology cannot handle interactions between multiple predictors, which limits the accuracy of predicting cases with atypical injuries. To approach this problem, we have proposed a new method for developing a model capable of handling the interactions in cases with multiple injuries.

(2) The proposed method of Bayesian averaging over DT models has outperformed the standard TRISS in terms of accuracy of predictions and has

provided reliable estimates of the predictive posterior distribution for patients who obtained multiple injuries.

(3) The Maximum a Posterior DT model has been found and described for purposes of interpenetration of predictions.

(4) A web-based application for trauma care practitioners has been developed and made available for evaluation.

4.2. Weaknesses

As discussed in Section 1, the accuracy of the TRISS model is acceptable when the types and severities of patient injuries are typical, and the number of most severe injuries is up to three. For patients with a larger number of injuries, the accuracy has to be improved. Along with other reasons, this motivated us to develop a new model capable of predicting outcomes of patients with multiple injuries more accurately.

To achieve this aim, the new model was developed on records of patients in Group 1, as described in Section 2. These patients were registered with four and more injuries, and so the mortality in this group was 7.78%. Mortality of patients in Group 2, who obtained up to three injuries, was 2.69% which is significantly lower.

Within this approach, it is not surprising to see that the developed model outperforms the TRISS model for patients at a high risk of death, and *vice versa* the TRISS outperforms the new model for patients at a low risk. The calibration curves calculated for the TRISS and proposed BDT models plotted on Fig. 1 show that the TRISS model has a better calibration for the predicted probabilities between 0 to 0.1. The TRISS predicts mortality 0.0128 against the observed mortality 0.0102, having a difference of 0.0026. At the same time, the BDT model predicts a probability 0.0130 for the observed mortality 0.0093, having a higher difference of 0.0037. Thus in this patient group both models overestimate the risk of death, and BDT predicts a higher probability. This problem however cannot be resolved by direct mixing of cases from Groups 1 and 2 without affecting precision. Therefore ways of improving the accuracy of predicting patients at low risk have to be further investigated in future work.

4.3. Conclusion and future work

There exist unexplained deviations in the TRISS calibration curve, which affect accuracy and reliability of predictions. Trauma care practitioners have also found that the prediction of outcomes for patients with multiple injuries is not reliable and has to be improved.

For improving the accuracy and reliability of predictions we have proposed a Bayesian method which was compared with TRISS an a data set including 447,176 cases from the US NTDB, the main data repository in trauma care research. We compared the goodness-of-fit of the proposed and TRISS models and found a significant improvement. The proposed method has improved the prediction accuracy in terms of AUC ($p = 3.89 \cdot 10^{-18}$).

The proposed Bayesian method has been implemented as a web application to support trauma care practitioners. The web application is accessible from the

user's device. Further improvements of prediction accuracy could be achieved with new variables which can be added to the screening tests described in Section 2.1. An improvement of the proposed method can be focused on a group of patients with low mortality where risks are overestimated.

Appendix A. An MCMC sampler

In practice, the dimensionality of models is typically unknown and can largely vary, so the desired approximation of $p(\Theta|D)$ is achieved with the Reversible Jump (RJ) extension of MCMC [54]. DT models are grown on given data by the RJ MCMC sampler according to the strategy proposed in [42]. The sampler aims to search model parameters Θ by making the following types of moves:

1. *Birth*. To randomly split the data points falling in one of the terminal nodes by adding a new splitting node with a variable and rule drawn from a given prior.
2. *Death*. To randomly pick a DT splitting node with two terminal nodes to be assigned a single terminal node with the merged data points.
3. *Change-split*. To randomly pick a splitting node and assign it a new splitting variable and rule drawn from a given prior.
4. *Change-rule*. To randomly pick a splitting node and assign it a new rule drawn from a given prior.

Making the birth and death moves the sampler can change the dimensionality of Θ, and so these moves have to be reversible. The change moves are made in order to search the parameters Θ within the current dimensionality of the DT model.

Acknowledgements

This research was partly supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant GR/R24357/01 "Critical Systems and Data-Driven Technology".

References

[1] T. Bailey, R. Everson, J. Fieldsend, W. Krzanowski, D. Partridge, V. Schetinin, Representing classifier confidence in the safety critical domain – an illustration from mortality prediction in trauma cases, Neural Computing and Applications 16 (3) (2007) 1–10. doi:10.1007/s00521-006-0053-y.

[2] H.-Y. Lu, T.-C. Li, Y.-K. Tu, J.-C. Tsai, H.-S. Lai, L.-T. Kuo, Predicting long-term outcome after traumatic brain injury using repeated measurements of glasgow coma scale and data mining methods., J. Medical Systems 39 (2015) 14. doi:10.1007/s10916-014-0187-x.

[3] J. González-Robledo, F. Martín-González, M. Sánchez-Barba, F. Sánchez-Hernández, M. N. Moreno-García, Multiclassifier systems for predicting neurological outcome of patients with severe trauma and polytrauma in intensive care units, Journal of Medical Systems 41 (9) (2017) 136. doi:10.1007/s10916-017-0789-1.

[4] B. M. Patil, R. C. Joshi, D. Toshniwal, S. Biradar, A new approach: Role of data mining in prediction of survival of burn patients, Journal of Medical Systems 35 (6) (2011) 1531–1542. doi:10.1007/s10916-010-9430-2.

[5] C. R. Boyd, M. A. Tolson, W. S. Copes, Evaluating trauma care: The TRISS method, Journal of Trauma 27 (1984) 370–378.

[6] P. Kilgo, J. Meredith, T. Osler, Injury severity scoring and outcomes research, in: D. V. Feliciano, K. L. Mattox, E. E. Moore (Eds.), Trauma (6th ed), New York, McGraw-Hill, 2008, pp. 223–230.

[7] O. Bouamra, A. Wrotchford, S. Hollis, A. Vail, M. Woodford, F. Lecky, A new approach to outcome prediction in trauma: A comparison with the TRISS model, Journal of Trauma 61 (3) (2006) 701–710. doi:10.1097/01.ta.0000197175.91116.10.

[8] R. Lefering, S. Huber-Wagner, U. Nienaber, M. Maegele, B. Bouillon, Update of the trauma risk adjustment model of the traumaregister dgu™: the revised injury severity classification, version ii, Critical Care 18 (5) (2014) 476. doi:10.1186/s13054-014-0476-2.

[9] B. J. Gabbe, P. A. Cameron, R. Wolfe, TRISS: Does It Get Better than This?, Academic Emergency Medicine 11 (2) (2004) 181–186. doi:10.1197/j.aem.2003.08.019.

[10] P. J. Schluter, A. Nathens, M. L. Neal, S. Goble, C. M. Cameron, T. M. Davey, R. J. McClure, Trauma and injury severity score (TRISS) coefficients 2009 revision, Journal of Trauma-Injury Infection & Critical Care 68 (4) (2017) 761–770. doi:10.1097/TA.0b013e3181d3223b.

[11] F. Rogers, T. Osler, M. Krasne, A. Rogers, E. Bradburn, J. Lee, D. Wu, N. McWilliams, M. Horst, Has TRISS become an anachronism? A comparison of mortality between the National Trauma Data Bank and major trauma outcome study databases, Journal of Trauma and Acute Care Surgery 73 (2) (2012) 326–331. doi:10.1097/TA.0b013e31825a7758.

[12] D. W. Hosmer, T. Hosmer, S. Le Cessie, S. Lemeshow, A comparison of goodness-of-fit tests for the logistic regression model, Statistics in Medicine 16 (9) (1997) 965–980.

[13] E. Steyerberg, A. Vickers, N. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. Pencina, M. Kattan, Assessing the performance of prediction models: A framework for traditional and novel measures, Epidemiology 21 (1) (2010) 128–138. doi:10.1097/EDE.0b013e3181c30fb2.

[14] D. Becalick, T. Coats, Comparison of artificial intelligence techniques with UKTRISS for estimating probability of survival after trauma. UK Trauma and Injury Severity Score, Journal of Trauma 51 (1) (2001) 123–133.

[15] P. Magni, G. Sparacino, R. Bellazzi, G. M. Toffolo, C. Cobelli, Insulin minimal model indexes and secretion: Proper handling of uncertainty by a Bayesian approach, Annals of Biomedical Engineering 32 (7) (2004) 1027–1037. doi:10.1023/B:ABME.0000032465.75888.91.

[16] W. J. Krzanowski, T. C. Bailey, D. Partridge, J. E. Fieldsend, R. M. Everson, V. Schetinin, Confidence in classification: A Bayesian approach, Journal of Classification 23 (2) (2006) 199–220. doi:10.1007/s00357-006-0013-3.

[17] A. Achilleos, C. Loizides, M. Hadjiandreou, T. Stylianopoulos, G. D. Mitsis, Multiprocess dynamic modeling of tumor evolution with Bayesian tumor-specific predictions, Annals of Biomedical Engineering 42 (5) (2014) 1095–1111. doi:10.1007/s10439-014-0975-y.

[18] D. Denison, C. Holmes, B. Mallick, A. Smith, Bayesian Methods for Nonlinear Classification and Regression, Wiley, 2002.

[19] M. A. Negrin, J. Nam, A. H. Briggs, Bayesian solutions for handling uncertainty in survival extrapolation, Medical Decision Making 37 (4) (2016) 367–376. doi:10.1177/0272989X16650669.

[20] V. Schetinin, C. Maple, A Bayesian model averaging methodology for detecting EEG artifacts, in: 2007 15th International Conference on Digital Signal Processing, Cardiff, 2007, pp. 499–502. doi:10.1109/ICDSP.2007.4288628.

[21] L. Jakaite, V. Schetinin, Feature selection for Bayesian evaluation of trauma death risk, in: The 14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics, Springer, 2008, pp. 123–126.

[22] V. Schetinin, L. Jakaite, Extraction of features from sleep eeg for Bayesian assessment of brain development, PLoS ONE 2 (3). doi:10.1371/journal.pone.0174027.

[23] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman and Hall, 1984.

[24] C. Robert, G. Casella, Monte Carlo Statistical Methods, Springer Texts in Statistics, Springer, 2004.

[25] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning, The MIT Press, 2009.

[26] V. Schetinin, L. Jakaite, J. Jakaitis, W. Krzanowski, Bayesian decision trees for predicting survival of patients: a study on the US National Trauma Data Bank, Computer Methods and Programs in Biomedicine 111 (3). doi:10.1016/j.cmpb.2013.05.015.

[27] V. Schetinin, L. Jakaite, W. J. Krzanowski, Prediction of survival probabilities with Bayesian decision trees, Expert Systems with Applications 40 (14) (2013) 5466 – 5476. doi:10.1016/j.eswa.2013.04.009.

[28] V. Schetinin, L. Jakaite, W. Krzanowski, Bayesian averaging over decision tree models for trauma severity scoring, Artificial Intelligence in Medicine (2017) –doi:10.1016/j.artmed.2017.12.003.

[29] W. J. Krzanowski, D. J. Hand, ROC Curves for Continuous Data, 1st Edition, Chapman & Hall/CRC, 2009.

[30] B. Ozenne, F. Subtil, D. Maucort-Boulch, The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases, Journal of Clinical Epidemiology 68 (8) (2015) 855–859. doi:10.1016/j.jclinepi.2015.02.010.

[31] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLOS ONE 10 (3) (2015) 1–21. doi:10.1371/journal.pone.0118432.

[32] The American College of Surgeons, National Trauma Data Bank, accessed: 04/01/2018 (2014).
URL http://www.facs.org/quality-programs/trauma/ntdb

[33] TraumaCalc: Bayesian prediction of trauma survival, accessed: 04/01/2018 (2016).
URL http://www.traumacalc.org/traumacalc/

[34] The American Association for the Surgery of Trauma, Injury scoring scale: A resource for trauma care professionals, accessed: 22/10/2017.
URL http://www.aast.org

[35] T. A. Gennarelli, E. Wodzin, Ais 2005: A contemporary injury scale, Injury 37 (12) (2006) 1083 – 1091, special Issue: Trauma Outcomes. doi:10.1016/j.injury.2006.07.009.

[36] J. Peng, K. Wheeler, J. Shi, J. I. Groner, K. J. Haley, H. Xiang, Trauma with injury severity score of 75: Are these unsurvivable injuries?, PLOS ONE 10 (7) (2015) 1–11. doi:10.1371/journal.pone.0134821.

[37] H. R. Champion, W. J. Sacco, W. S. Copes, D. S. Gann, T. a. Gennarelli, M. E. Flanagan, A revision of the Trauma Score, The Journal of trauma 29 (5) (1989) 623–629. doi:10.1097/00005373-198905000-00017.

[38] T. Osler, L. Glance, J. Buzas, D. Mukamel, J. Wagner, A. Dick, A trauma mortality prediction model based on the anatomic injury scale, Annals of Surgery 247 (6) (2008) 1041–1048. doi:10.1097/SLA.0b013e31816ffb3f.

[39] V. Schetinin, J. Schult, Learning polynomial networks for classification of clinical electroencephalograms, Soft Computing 10 (4) (2006) 397–403. doi:10.1007/s00500-005-0499-3.

[40] E. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, Statistics for Biology and Health, Springer New York, 2010.

[41] G. Meyfroidt, F. Güiza, D. Cottem, W. De Becker, K. Van Loon, J.-M. Aerts, D. Berckmans, J. Ramon, M. Bruynooghe, G. Van den Berghe, Computerized prediction of intensive care unit discharge after cardiac surgery: development and validation of a Gaussian processes model, BMC Medical Informatics and Decision Making 11 (1) (2011) 64. doi:10.1186/1472-6947-11-64.

[42] V. Schetinin, J. E. Fieldsend, D. Partridge, W. J. Krzanowski, R. M. Everson, T. C. Bailey, A. Hernandez, Comparison of the Bayesian and randomized decision tree ensembles within an uncertainty envelope technique, Journal of Mathematical Modelling and Algorithms 5 (4) (2006) 397–416. doi:10.1007/s10852-005-9019-9.

[43] K. Brohi, TRISS - Overview and desktop calculator, accessed: 04/01/2018 (2012).
URL http://www.trauma.org/index.php/main/article/387

[44] H. Chipman, E. George, R. McCullock, Bayesian CART model search, Journal of American Statistics 93 (1998) 935–960.

[45] L. Jakaite, V. Schetinin, C. Maple, Bayesian assessment of newborn brain maturity from two-channel sleep electroencephalograms, Computational and Mathematical Methods in Medicine 2012 (2012) 1–7. doi:10.1155/2012/629654.

[46] V. Schetinin, L. Jakaite, Classification of newborn EEG maturity with Bayesian averaging over decision trees, Expert Systems with Applications 39 (10) (2012) 9340–9347.

[47] A. A. Kramer, J. E. Zimmerman, Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited., Critical Care Medicine 35 (9) (2007) 2052–2056.

[48] M. Pennell, A. Bartley, S. Lemeshow, G. Phillips, A strategy for evaluating goodness-of-fit for a logistic regression model using the hosmer-lemeshow test on samples from a large data set, in: JSM Proceedings, Section on Statistics in Epidemiology, American Statistical Association, 2017.

[49] N. R. Cook, J. E. Buring, P. Ridker, The effect of including c-reactive protein in cardiovascular risk prediction models for women, Annals of Internal Medicine 145 (1) (2006) 21–29. doi:10.7326/0003-4819-145-1-200607040-00128.

[50] P. Paul, M. L. Pennell, S. Lemeshow, Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets, Statistics in Medicine 32 (1) (2013) 67–80. doi:10.1002/sim.5525.

[51] Philip R. Lee Institute for Health Policy Studies, Summary of NQF-endorsed intensive care outcomes models for risk adjusted mortality and length of stay, http://healthpolicy.ucsf.edu/icu-outcomes, accessed: 22/10/2017.

[52] T. Toma, A. Abu-Hanna, R.-J. Bosman, Discovery and inclusion of SOFA score episodes in mortality prediction, Journal of Biomedical Informatics 40 (6) (2007) 649 – 660. doi:10.1016/j.jbi.2007.03.007.

[53] D. Benjamin, D. R. Mandel, J. Kimmelman, Can cancer researchers accurately judge whether preclinical reports will reproduce?, PLoS Biology 15 (6) (2017) 1–17. doi:10.1371/journal.pbio.2002212.

[54] P. J. Green, Reversible jump Markov chain Monte Carlo and Bayesian model determination, Biometrika 82 (1995) 711–732.